Multilingual NMT with a language-independent attention bridge

Vazquez Carrillo, Juan Raul

The Association for Computational Linguistics
2019

Vazquez Carrillo , J R , Raganato , A , Tiedemann , J & Creutz , M 2019 , Multilingual NMT with a language-independent attention bridge . in I Augenstein , S Gella , S Ruder , K Kann , B Can , J Welbl , A Conneau , X Ren & M Rei (eds) , The 4th Workshop on Representation Learning for NLP (RepL4NLP-2019) : Proceedings of the Workshop . The Association for Computational Linguistics , Stroudsburg , pp. 33-39 , Workshop on Representation Learning for NLP , Florence , Italy , 02/08/2019 .

http://hdl.handle.net/10138/304660

Multilingual NMT with a language-independent attention bridge

Raúl Vázquez,* **Alessandro Raganato,*†** **Jörg Tiedemann*** **and Mathias Creutz***
*University of Helsinki, Department of Digital Humanities
†Basement AI
{name.surname}@helsinki.fi

Abstract

In this paper, we propose an architecture for machine translation (MT) capable of obtaining multilingual sentence representations by incorporating an intermediate *attention bridge* that is shared across all languages. We train the model with language-specific encoders and decoders that are connected through an inner-attention layer on the encoder side. The attention bridge exploits the semantics from each language for translation and develops into a language-agnostic meaning representation that can efficiently be used for transfer learning. We present a new framework for the efficient development of multilingual neural machine translation (NMT) using this model and scheduled training. We have tested the approach in a systematic way with a multi-parallel data set. The model achieves substantial improvements over strong bilingual models and performs well for zero-shot translation, which demonstrates its ability of abstraction and transfer learning.

1 Introduction

Neural machine translation (NMT) provides an ideal setting for multilingual MT because it can efficiently share model parameters and take advantage of the various similarities found by the model in the hidden layers and word embeddings (Firat et al., 2016a; Johnson et al., 2017; Blackwood et al., 2018). Furthermore, multilingual NMT has the potential of considerably improving the performance of neural translation systems for low-resource languages (Lakew et al., 2017) and enables zero-shot translation, i.e., translating between language pairs that were not seen during training (Firat et al., 2016b; Johnson et al., 2017).

For this study we focus on models for multilingual translation that learn language-agnostic representations, where we outline the development of a language-independent representation based on an *attention bridge* shared across all languages. For this, we apply an architecture based on shared self-attention with language-specific encoders and decoders that can easily scale to a large number of languages while addressing the task of obtaining language-independent sentence embeddings (Cífka and Bojar, 2018; Lu et al., 2018; Lin et al., 2017). Those embeddings are created from the encoder's self-attention and connect to the language-specific decoders that attend to them, hence the name 'bridge'. Additionally, we add a penalty term to avoid redundancy in the shared layer. More details of the architecture are given in section 2.

To summarise our contributions, we **i)** present a multilingual translation system that efficiently tackles the task of learning language-agnostic sentence representations; **ii)** verify that this model enables effective transfer learning and zero-shot translation through the shared representation layer; and **iii)** show that multilingually trained embeddings improve the majority of downstream and sentence probing tasks demonstrating the abstractions learned from the combined translation tasks.

2 Model Architecture

Our architecture follows the standard setup of an encoder-decoder model of machine translation with a traditional attention mechanism (Bahdanau et al., 2015; Luong et al., 2015). However, to enable multilingual training we augment the network with language-specific encoders and decoders trainable with a language-rotating scheduler (Dong et al., 2015; Schwenk and Douze, 2017). We also incorporate a self-attention layer (attention bridge), shared among all language pairs, to serve as a language-agnostic layer (Cífka and Bojar, 2018; Lu et al., 2018)

Attention bridge: Each encoder takes as input

Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pages 33–39
Florence, Italy, August 2, 2019.

a sequence of tokens $(x_1, \ldots, x_n)$ and produces n d_h–dimensional hidden states, $H = (h_1, \ldots, h_n)$ with $h_i \in \mathbb{R}^{d_h}$, in our case using a bidirectional long short-term memory (LSTM) (Graves and Schmidhuber, 2005)[1]. Next, we encode this variable length sentence-embedding matrix H into a fixed size $M \in \mathbb{R}^{d_h \times k}$ capable of focusing on k different components of the sentence (Lin et al., 2017; Chen et al., 2018; Cífka and Bojar, 2018), using self-attention as follows:

$$A = softmax\left(W_2 \text{ReLU}(W_1 H^T)\right) \quad (1)$$

$$M = AH \quad (2)$$

where $W_1 \in \mathbb{R}^{d_w \times d_h}$ and $W_2 \in \mathbb{R}^{k \times d_w}$ are weight matrices, with d_w a hyper-parameter set arbitrarily, and k is the number of *attention heads* in the attention bridge.

Each decoder follows a common attention mechanism in NMT (Luong et al., 2015), with an initial state computed by mean pooling over M, and using M instead of the hidden states of the encoder for computing the context vector.

Penalty term: The attention bridge matrix M from Eq. (2) could learn repetitive information for different attention heads. To address this issue, we add a penalty term to the loss function, proven effective in related work (Lin et al., 2017; Chen et al., 2018; Tao et al., 2018), which forces each vector to focus on different aspects of the sentence by making the columns of A to be approximately orthogonal in the Frobenius norm:

$$\mathcal{L} = -\log\left(p\left(Y|X\right)\right) + \left\|AA^T - I\right\|_F^2, \quad (3)$$

where the Frobenius norm of a matrix A can be defined as the sum of the squared singular values of A. By incorporating this term into the loss function we force matrix AA^T to be similar to the identity matrix, that is, $\sum_j a_{ij} a_{ji} \approx 1$. Additionally, considering the fact that the rows of A sum to 1, with entries in $[0, 1]$, it follows that the columns of A will be forced to be approximately orthogonal, and hence penalize redundancy, similar to the double stochastic attention in Xu et al. (2015).

[1]Note that the attention bridge is independent of the underlying encoder and decoder (Lu et al., 2018). While we use a biLSTM, it could be replaced with a gated recurrent unit (GRU) (Cho et al., 2014), a transformer type network (Vaswani et al., 2017) or with a convolutional neural network (CNN) (Gehring et al., 2017).

3 Experimental Setup

We conducted four translation experiments and tested the learned sentence representations via downstream tasks. We used the multi30k dataset (Elliott et al., 2016) for training and validation in all available languages: Czech, German, French and English, and tested the trained model with the flickr 2016 test data of the same dataset and obtained BLEU scores using the sacreBLEU script[2] (Post, 2018). We lowercased, normalized and tokenized using the Moses toolkit (Koehn et al., 2007), and applied a 10K-operations Byte Pair Encoding (BPE) model per language (Sennrich et al., 2016).

Each encoder consists of 2 stacked BiLSTMs of size $d_h = 512$, i.e., the hidden states per direction are of size 256. Each decoder includes 2 stacked unidirectional LSTMs with hidden states of size 512. For the model input and output, the word embeddings have dimension $d_x = d_y = 512$. We used an attention bridge layer with 10 attention heads with $d_w = 1024$, the dimensions of W_1 and W_2 from Eq. (1). We chose $k = 10$ because the mean length of a preprocessed sentence in the training data is 13.2 tokens in our case. Choosing a much smaller k would create a bottleneck in the flow of information, and a bigger one would make the model slower and prone to overfitting (Raganato et al., 2019).

We used a Stochastic Gradient Descent (SGD) optimizer with a learning rate of 1.0 and batch size 64, and selected the best model on the development set for each experiment. We implemented our model on top of an OpenNMT-py (Klein et al., 2017) fork, which we make available for reproducibility purposes.[3]

4 Results

First, we verify the correct functionality of the architecture in a bilingual setting, which will become our baseline for comparison to the multilingual models - both with and without an attention bridge.

On the *left* side of Table 1, we can see that the attention bridge model is almost on par with the standard bilingual model for all language pairs in our data set. A decrease in performance is to be

[2]with signature BLEU+case.lc+numrefs.1+smooth.exp+tok.13a+version.1.2.11

[3]https://github.com/Helsinki-NLP/OpenNMT-py/tree/att-brg.

BILINGUAL					{DE,FR,CS} ↔ EN				M-2-M			
src/tgt	EN	DE	CS	FR	EN	DE	CS	FR	EN	DE	CS	FR
EN	-	36.78	28.00	55.96	-	37.85	29.51	57.87	-	37.70	29.67	55.78
DE	39.00	-	23.44	38.22	39.39	-	*0.35*	*0.83*	40.68	-	26.78	41.07
CS	35.89	28.98	-	36.44	37.20	*0.65*	-	*1.02*	38.42	31.07	-	40.27
FR	49.54	32.92	25.98	-	48.49	*0.60*	*0.30*	-	49.92	34.63	26.92	-
BILINGUAL + ATT BRIDGE					{DE,FR,CS} ↔ EN + MONOLING				M-2-M + MONOLINGUAL			
	EN	DE	CS	FR	EN	DE	CS	FR	EN	DE	CS	FR
EN	-	35.85	27.10	53.03	-	38.92	30.27	57.87	-	38.48	30.47	57.35
DE	38.19	-	23.97	37.40	40.17	-	*19.50*	*26.46*	41.82	-	26.90	41.49
CS	36.41	27.28	-	36.41	37.30	*22.13*	-	*22.80*	39.58	31.51	-	40.87
FR	48.93	31.70	25.96	-	50.41	*25.96*	*20.09*	-	50.94	35.25	28.80	-

Table 1: BLEU scores obtained in the experiments. *Left:* Bilingual models, our baselines. *Center:* Models trained on {De,Fr,Cs}↔En, with zero-shot translations in italics. *Right:* Many-to-many model. Both zero-shot and M-2-M translations improve significantly when including monolingual data. (Best results in green cells.)

expected since we pass the information through a fixed size representation made out of 10 self-attention heads without including multilingual information. However, the drop is less than one BLEU point except for English to French, which seems to be an exceptional outlier.

With this result we can justify the validity of the architecture assuring that the additional bottleneck does not create significant deterioration and we can move on with the multilingual models.

4.1 Many-To-One and One-To-Many Models

The power of the attention bridge comes from its ability to share information across various language pairs. We now assess the effects of multilingual information on the translation of individual language pairs, by training many-to-one and one-to-many models. This setup allows us to test the abstraction potential of the attention bridge and its effectiveness to encode multilingual information in zero-shot translation.

First we trained a {De,Fr,Cs}↔En model (Table 1 *(center-top)*), which resulted in substantial improvements for the language pairs seen during training, exceeding both bilingual baselines. However, this model is entirely incapable of performing zero-shot translations. We believe that the inability of the model to generalize to unseen language-pairs arises from the fact that every non-English encoder (or decoder) only learned to process information that was to be decoded into English (or encoded from English input), a finding consistent with Lu et al. (2018). To address this problem, we incorporate monolingual data during training, that is, for each available language A, we included pairs of identical copies of each sentence in A in the training data. All examples come from the same parallel corpus as before and no additional data is used.

As a consequence, we see a remarkable increase in the BLEU scores, including a substantial boost for the language pairs not seen during training (Table 1 *(center-bottom)*). It seems that the monolingual data informs the model that English is not the unique source/target language. Additionally, there is a positive effect on the seen language pairs (up to almost 2 BLEU points for French to English), the cause of which is not immediately evident. It is possible that the shared layer acquires additional information that can be included in the abstraction process yet not available to the other models.

4.2 Many-to-Many Models

We also tested the architecture in a many-to-many setting with all language pairs included, and summarize our results in Table 1 *(right)*. As in the previous case, we compare settings with and without monolingual training data.

The inclusion of language pairs results in an improved performance when compared to the bilingual baselines, as well as the Many↔En cases, except for the En→Fr and En→De tasks. Moreover, the addition of monolingual data leads to even higher scores, producing the overall best model. The improvements in BLEU range from 1.40 to a remarkable 4.43 when compared to the standard bilingual model.

The zero-shot translation capabilities also deserve a closer look. Figure 1 summarizes a systematic evaluation in which we trained six different models where we include all but one of the available language pairs in training. The cyan bars illustrate the performance of the model on the unseen language pairs compared to our best multi-

lingual model (in red) and the bilingual, fully supervised model (in dark blue). Note, that those zero-shot models are generally better than the ones from the previously discussed {De,Fr,Cs,En}↔En model in Table 1. In most cases, they come very close to the supervised model and even fare well against the multilingual ones.

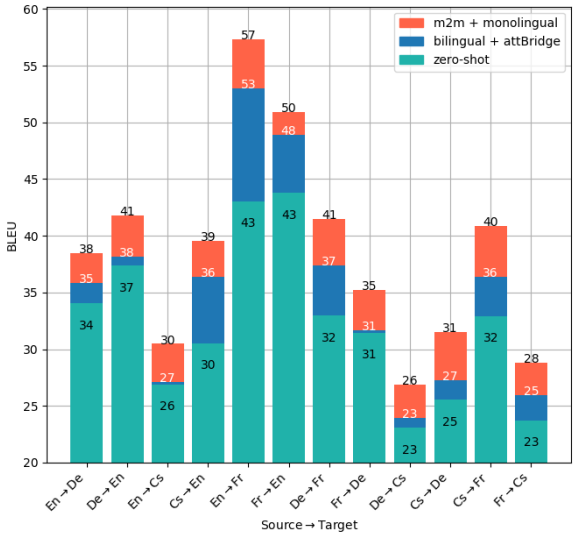

Figure 1: For every language pair, we compare the BLEU scores between our best model (M-2-M with monolingual data), the zero-shot of the model trained without that specific language pair and the bilingual model of that language pair.

5 Downstream Tasks

We apply the sentence representations learned by our model to downstream tasks collected in the SentEval toolkit (Conneau and Kiela, 2018) to evaluate the quality of our language-agnostic sentence embeddings. We run each experiment with five different seeds, and present the average of these scores in Table 2, where we compare our bilingual models against a baseline consisting of the best score achieved by the bilingual models with attention bridge. Since our models were trained on limited data and are not directly comparable to models trained on large-scale data sets, for comparison purposes we present results obtained with GloVe-BoW vectors (Pennington et al., 2014) trained with the same BPE-encoded data as the models.

The sentence embeddings produced by the multilingual models show consistent improvements, for the classification tasks of the SentEval collection, with only two exceptions. Moreover, our many-to-many model obtains better results in the SICK Relatedness (SICKR) and STS-Benchmark (STS-B); that is, the trainable semantic similarity tasks. [2]

For the SentEval probing tasks (Conneau et al., 2018) we use the default recommended settings, i.e., a multilayer perceptron classifier with sigmoid nonlinearity, 200 hidden units, and 0.1 dropout rate. Again, we can observe improvements in the majority of cases when adding multiple languages to the training procedure. Remarkably, we observe a significant increment on the accuracy for the specific tasks of Length (superficial property), Top Constituents (syntactic property) and Object Number (semantic information) when training the encoders with multilingual data. Multilingual models outperform the bilingual models in all but one test.

Task	Baseline	M ↔ EN	M-2-M	GloVe-BoW
Downstream Tasks				
CR	68.52	68.32	69.01	63.97
MR	60.08	60.40	61.80	52.32
MPQA	73.51	72.98	73.28	68.76
SUBJ	77.25	78.64	80.88	58.75
SST2	61.92	62.02	62.24	54.68
SST5	31.15	32.10	31.83	28.20
TREC	67.75	69.84	66.40	21.16
MRPC	70.96	68.83	70.43	64.87
SNLI	61.75	64.52	65.12	35.05
SICKE	74.85	75.46	76.92	56.62
SICKR	*0.652*	*0.659*	*0.677*	*0.174*
STS-B	*0.616*	*0.618*	*0.630*	*0.163*
Probing Tasks				
Length	80.76	84.76	85.41	30.90
WC	10.02	9.56	9.13	0.22
Depth	32.14	33.05	31.60	20.66
TopConst	40.12	44.04	39.76	11.48
BShift	57.41	58.35	59.76	50.08
Tense	67.61	69.36	68.27	54.72
SubjNum	68.55	69.67	69.89	54.32
ObjNum	70.01	72.19	73.29	60.58
SOMO	49.90	49.46	50.12	50.03
CoordInv	61.38	60.57	62.21	49.88

Table 2: Scores obtained in the SentEval tasks. The BASELINE column reports the best score among the bilingual models + att bridge. Green cells indicate the highest score. All tasks show the accuracy of the model except for SICKR and STS-B tasks, which include Pearson mean values.

[2] However, the non-trainable semantic similarity tasks exhibited decreasing scores for multilingual models (not shown here due to space limitations). This can be explained by the fact that the additional information encoded in our multilingual embeddings cannot effectively be separated from the information that is necessary for monolingual similarity measures, without further training.

6 Effect of the Penalty Term

In order to study the effect of the penalty term, we train additional bilingual models, without using the penalty term (Eq. 3) in the training. We then compare BLEU scores, where the penalty term is present and absent, as shown in Table 3. Overall, both types of models show performance in the same ballpark yielding similar results. As discussed in Lin et al. (2017), the quantitative effect of the penalty term might not be obvious for some tasks, while keeping the positive effect of encouraging the attentive matrix to be focused on different aspects of the sentence.

	WITH PENALTY TERM			
	EN	DE	CS	FR
EN	-	35.85	27.10	53.03
DE	38.19	-	23.97	37.40
CS	36.41	27.28	-	36.41
FR	48.93	31.70	25.96	-
	WITHOUT PENALTY TERM			
	EN	DE	CS	FR
EN	-	34.67	27.22	54.39
DE	38.70	-	23.44	38.2
CS	35.76	28.50	-	36.4
FR	48.76	31.60	25.55	-

Table 3: BLEU scores obtained with the BILINGUAL + ATT BRIDGE models in the experiments with and without penalty term.

While the effect of the penalty term might not be very significant in this case, we note that adding the penalty term does not hurt the performance while helping the model not to learn potential redundant information.

7 Conclusion

We propose a multilingual NMT architecture with three modifications to the common attentive encoder-decoder architecture: language-specific encoders and decoders, a shared language-independent attention bridge and a penalty term that forces this layer to attend different parts of the input sentence. This constitutes a multilingual translation system that efficiently incorporates transfer learning and can also tackle the task of learning multilingual sentence representations. The results suggest that the attention bridge layer can efficiently share parameters in a multilingual setting, increasing up to 4.4 BLEU points compared to the baselines. Additionally, we make use of the sentence representations produced by the shared attention bridge of the trained models for downstream-testing, which helped us to verify the generalization capabilities of the model. The results suggest that sentence embeddings improve with additional languages involved in training the underlying machine translation model.

Acknowledgments

This work is part of the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 771113).

The authors gratefully acknowledge the support of the Academy of Finland through project 314062 from the ICT 2023 call on Computation, Machine Learning and Artificial Intelligence. We thank the participants that contributed to the project we lead during the 13th MT Marathon in Prague. We are particularly grateful with Chris Hokamp, whose help was crucial during that time. Finally, We would also like to acknowledge NVIDIA and their GPU grant.

References

Dzimitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ICLR 2015*.

Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual Neural Machine Translation with Task-Specific Attention. *Proceedings of the 27th Conference on Computational Linguistics*, pages 3112–3122.

Qian Chen, Zhen-Hua Ling, and Xiaodan Zhu. 2018. Enhancing sentence embedding with generalized pooling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1815–1826.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Ondřej Cífka and Ondřej Bojar. 2018. Are bleu and meaning representation in opposition? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1362–1371.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-Task Learning for Multiple Language Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

D. Elliott, S. Frank, K. Sima'an, and L. Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016b. Zero-Resource Translation with Multi-Lingual Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *Proc. of ICML*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Googles multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, and Alexandra Constantinand Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. *Annual Meeting of the Association for Computational Linguistics (ACL)*, Demo and Poster Sessions.

Surafel M Lakew, Quintino F Lotito, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Improving Zero-Shot Translation of Low-Resource Languages. *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 113–119.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *5th International Conference on Learning Representations (ICLR 2017)*.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Alessandro Raganato, Raúl Vázquez, Mathias Creutz, and Jörg Tiedemann. 2019. An evaluation of language-agnostic inner-attention-based representations in machine translation. In *Proceedings of The Fourth Workshop on Representation Learning for NLP (RepL4NLP)*. Association for Computational Linguistics.

Holger Schwenk and Matthijs Douze. 2017. Learning Multilingual Sentence Representations with Neural Machine Translation. In *2nd Workshop on Representation Learning for NLP*, pages 157–167. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words

with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get The Point of My Utterance! Learning Towards Effective Responses with Multi-Head Attention Mechanism. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4418–4424, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image captiongeneration with visual attention. *Proceedings of the 32nd International Conference on Machine Learning*, pages 1532–1543.